Robust Variable Selection in Discriminant Analysis

Stefan Van Aelst and Gert Willems

Abstract We consider robust linear discriminant rules that are obtained by replacing the empirical means and covariance in the classical discriminant rules by robust S or MM-estimates of location and scatter. We consider the problem of selecting the variables that are relevant for separating the groups. We propose to use a fast and robust bootstrap method to test which variables contribute significantly to the discrimination of the classes. Using the fast and robust bootstrap is necessary because classical bootstrap methods may be unstable as well as extremely time-consuming when robust estimates are involved. The fast and robust bootstrap test can be used as an inclusion/exclusion criterion in step-by-step algorithm that aim to select the set of predictors that yields the best discriminatory power.

Key words: bootstrap, classification and pattern recognition, robustness, variable selection

1 Introduction

Linear discriminant rules are widely used to optimally separate multivariate observations in two or more populations. For simplicity, in this paper, we consider situations with two p-dimensional populations, Π_1 and Π_2 with corresponding population means μ_1 and μ_2. Moreover, we assume that the two populations have a common covariance matrix Σ and have equal prior probabilities. The linear Bayes rule then classifies an observation $\mathbf{x} \in \mathbb{R}^p$ into population Π_1 if $d_1^L(\mathbf{x}) > d_2^L(\mathbf{x})$, where

Stefan Van Aelst
Ghent University, Department of Applied Mathematics and Computer Science, Krijgslaan 281 S9, B-9000 Gent, Belgium e-mail: Stefan.VanAelst@UGent.be

Gert Willems
Ghent University, Department of Applied Mathematics and Computer Science, Krijgslaan 281 S9, B-9000 Gent, Belgium e-mail: Gert llWillems@gmail.com

$$d_j^L(\mathbf{x}) = \mu_j^t \Sigma^{-1} \mathbf{x} - \frac{1}{2} \mu_j^t \Sigma^{-1} \mu_j;\ j = 1,2, \tag{1}$$

and into population Π_2 otherwise. The direction $\mathbf{a}$ that best separates the two populations is then given by $(\mu_1 - \mu_2)\Sigma^{-1}$. The corresponding projection $\mathbf{a}^t\mathbf{x}$ is called the canonical variate or discriminant coordinate.

Since μ_1, μ_2 and Σ are unknown, they need to be estimated from an available training sample of the form $\mathscr{Z}_n = \{\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1}, \mathbf{x}_{21}, \ldots, \mathbf{x}_{2n_2}\} \subset \mathbb{R}^p$ with n_1 the number of samples from population 1, n_2 the number of samples from population 2 and $n = n_1 + n_2$, the total sample size. Fisher's classical linear discriminant analysis is based on the empirical means and covariances of the training data $\mathscr{Z}_n$. Robust linear discriminant analysis methods can be obtained by using robust estimates of the two centers and common scatter matrix (see e.g. Croux et al. (2008) and Bianco et al. (2008) and references therein).

Many robust estimators of multivariate location and scatter have been proposed in the literature, see e.g. Hubert et al. (2008) for a recent overview. In this paper we use the classes of S-estimators and MM-estimators (Tatsuoka and Tyler 2000) to robustly estimate the centers of the populations and their common scatter matrix. Inference for these estimators can be derived from their asymptotic distribution. However, this asymptotic distribution is mainly known for elliptical model distributions, an assumption which is not appropriate in those cases where robust estimation is most recommended, i.e. for data with outliers. The bootstrap is a computer-intensive alternative that can be more reliable for smaller sample sizes and for larger deviations from the central model. Moreover, because the bootstrap estimates the sampling distribution of the estimators, it has applications beyond the standard inference procedures of calculating standard errors, confidence intervals or p-values for hypothesis tests. For example, bootstrap can also be used to investigate the stability of a discriminant analysis.

Applying the standard bootstrap on robust estimators poses both a computational issue and a robustness issue. Calculating robust estimates is complex and requires a high computation time, which makes it infeasible to obtain a large number of recalculated robust estimates in a reasonable amount of time. Because bootstrap samples are drawn with replacement, the amount of outliers varies between bootstrap samples and can exceed the breakdown point of the estimator in some samples. For multivariate S and MM-estimators, both issues can be solved at once by the fast and robust bootstrap (FRB), introduced by Salibian-Barrera and Zamar (2002) in the context of robust regression based on MM-estimators. See e.g. Van Aelst and Willems (2005), Salibian-Barrera et al. (2006, 2008), Salibian-Barrera and Van Aelst (2008) and Roelant et al. (2009) for recent application of the FRB. Here, we use the FRB to obtain many recalculations of the robust S or MM-estimates of the locations and common scatter matrix in a linear discriminant analysis. These FRB estimates can then be used for inference purposes. Moreover, we also use the FRB to investigate which variables contribute significantly to the discriminant analysis and to select the most relevant variables. In this way we can investigate which variables carry the most discriminatory power.

The rest of this paper is organized as follows. In Section 2 we review multivariate S and MM-estimators. Section 3 explains the fast and robust bootstrap in this setting. In Section 4 we explain some useful applications of the FRB for discriminant analysis and illustrate the method on some real data examples and Section 5 concludes.

2 Multivariate S and MM-estimators

In the multivariate one-sample setting S-estimators are defined as follows. Given a sample $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}\subset\mathbb{R}^p$ and a function $\rho_0:[0,\infty[\to[0,\infty[$ which is bounded, increasing and sufficiently smooth, the S-estimates of location and scatter $(\widetilde{\mu}_n, \widetilde{\Sigma}_n)$ minimize $|C|$ subject to

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left([(\mathbf{x}_i-T)^tC^{-1}(\mathbf{x}_i-T)]^{\frac{1}{2}}\right)=b \tag{2}$$

among all $T\in\mathbb{R}^p$ and $C\in\mathrm{PDS}(p)$. Here, PDS(p) denotes the set of positive definite symmetric $p\times p$ matrices and $|C|$ denotes the determinant of the square matrix C. In this paper, the loss function ρ_0 is taken from the common class of Tukey biweight functions, given by $\rho_c(t)=\min(t^2/2-t^4/(2c^2)+t^6/(6c^4),c^2/6)$. The constant b is usually chosen to ensure consistency of the S-estimator at the normal model. The constant c in the Tukey biweight loss function ρ_c can be tuned to achieve a desired degree of robustness, but at the same time affects the efficiency of the S-estimators. Therefore, highly robust S-estimators can have quite low efficiency at normal distributions (see e.g. Salibian-Barrera et al. 2006).

Multivariate (one-sample) MM-estimators of location and shape have been introduced by Tatsuoka and Tyler (2000) to obtain highly robust and highly efficient estimators. Let $\widetilde{\Sigma}_n$ be the S-estimate of scatter and denote $\hat{\sigma}_n:=|\widetilde{\Sigma}_n|^{1/(2p)}$ the corresponding S-estimate of multivariate scale. Let ρ_1 be a loss function from the same class as ρ_0. Then, the multivariate MM-estimates of location and shape $(\widehat{\mu}_n,\widehat{\Gamma}_n)$ minimize

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1\left([(\mathbf{x}_i-T)^tG^{-1}(\mathbf{x}_i-T)]^{\frac{1}{2}}/\hat{\sigma}_n\right)$$

among all $(T,G)\in\mathbb{R}^p\times\mathrm{PDS}(p)$ for which $|G|$=1. The corresponding MM-estimator for the scatter matrix is given by $\widehat{\Sigma}_n=\hat{\sigma}_n^2\widehat{\Gamma}_n$.

MM-estimates inherit the robustness of the initial S-estimate of multivariate scale as determined by the loss function ρ_0, while the loss function ρ_1 can be tuned to obtain a high efficiency, e.g. 95%, at the normal model.

In robust linear discriminant we need a robust estimate of the common covariance matrix Σ. Based on the robust scatter estimates $\widehat{\Sigma}_{1n_1}$ and $\widehat{\Sigma}_{2n_2}$ for the two groups we can calculate the pooled scatter estimate $\widehat{\Sigma}_n$ as

$$\widehat{\Sigma}_n = \frac{n_1\widehat{\Sigma}_{1n_1} + n_2\widehat{\Sigma}_{2n_2}}{n_1+n_2}.$$

See e.g. Croux et al. (2008), and Bianco et al. (2008) among others.

Alternatively, following He and Fung (2000) the definition of the S and MM-estimators can be adjusted. Simultaneous S-estimates of the two locations and the common scatter matrix can be defined as the solution $\widehat{\mu}_{1n}$, $\widehat{\mu}_{2n}$ and $\widehat{\Sigma}_n$ that minimizes $|C|$ subject to

$$\frac{1}{n_1+n_2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}\rho\left([(\mathbf{x}_{ji}-T_j)^t C^{-1}(\mathbf{x}_{ji}-T_j)]^{\frac{1}{2}}\right) = b \tag{3}$$

among all $T_1, T_2 \in \mathbb{R}^p$ and $C \in \mathrm{PDS}(p)$. Similarly, simultaneous MM-estimates for the two locations and common shape/scatter can be defined.

3 Fast and robust bootstrap

The fast and robust bootstrap procedure assumes that the robust estimates can be written as a solution of a set of sufficiently smooth fixed point equations. Both S and MM-estimates can be written as the solution of a set of smooth fixed point equations as follows. Let $\hat{\theta}_n$ be a vector that collects all parameter estimates of interest. In our case, for S-estimates $\hat{\theta}_n$ contains the two location estimates and the common scatter estimate in vectorized form. In case of MM-estimates, the vector $\hat{\theta}_n$ additionally contains the MM location and (vectorized) shape estimates. Then, the estimates can be written as the solution of

$$\hat{\theta}_n = \mathbf{g_n}(\hat{\theta}_n) \tag{4}$$

where the function $g_n : \mathbb{R}^d \to \mathbb{R}^d$ depends on the sample $\mathscr{Z}_n$. Given a bootstrap sample $\mathscr{Z}_n^*$, which means a sample of size n_1+n_2 drawn with replacement from $\mathscr{Z}_n$, the recalculated estimate $\hat{\theta}_n^*$ is the solution of the corresponding fixed point equation $\hat{\theta}_n^* = \mathbf{g}_n^*(\hat{\theta}_n^*)$, where the function $\mathbf{g}_n^*$ now depends on $\mathscr{Z}_n^*$. Instead of calculating $\hat{\theta}_n^*$, we consider the simple and fast approximation $\hat{\theta}_n^{1*} := \mathbf{g}_n^*(\hat{\theta}_n)$. Note that $\hat{\theta}_n^{1*}$ is only a one-step approximation for $\hat{\theta}_n^*$ starting from the initial value $\hat{\theta}_n$, obtained for the original sample.

However, since we are keeping the estimates $\hat{\theta}_n$ fixed when calculating the approximations $\hat{\theta}_n^{1*}$, these approximations will underestimate the variability of the estimator. To remedy this, a linear correction can be applied as follows. Consider a Taylor expansion about $\hat{\theta}_n$'s limiting value θ,

$$\hat{\theta}_n = \mathbf{g}_n(\theta) + \nabla\mathbf{g}_n(\theta)(\hat{\theta}_n - \theta) + R_n,$$

where R_n is the remainder term and $\nabla\mathbf{g}_n(.) \in \mathbb{R}^{m\times m}$ is the matrix of partial derivatives. Assuming that the remainder term is negligible, this can be rewritten as

$$\sqrt{n}(\hat{\theta}_n - \theta) \dot{\sim} [\mathbf{I} - \nabla \mathbf{g}_n(\theta)]^{-1} \sqrt{n}(\mathbf{g}_n(\theta) - \theta),$$

where $\dot{\sim}$ denotes that both sides have the same limiting distribution. Under weak regularity conditions we will have that $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \dot{\sim} \sqrt{n}(\hat{\theta}_n - \theta)$ and $\sqrt{n}(\mathbf{g}_n^*(\theta) - \theta) \dot{\sim} \sqrt{n}(\mathbf{g}_n(\hat{\theta}_n) - \hat{\theta}_n)$ (see e.g. Salibian-Barrera et al. (2008)). If we furthermore approximate $[\mathbf{I} - \nabla \mathbf{g}_n(\theta)]^{-1}$ by $[\mathbf{I} - \nabla \mathbf{g}_n(\hat{\theta}_n)]^{-1}$ we obtain

$$\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \dot{\sim} [\mathbf{I} - \nabla \mathbf{g}_n(\hat{\theta}_n)]^{-1} \sqrt{n}(\mathbf{g}_n^*(\hat{\theta}_n) - \hat{\theta}_n).$$

We now define the fast and robust bootstrap estimates as

$$\hat{\theta}_n^{R*} := \hat{\theta}_n + [\mathbf{I} - \nabla \mathbf{g}_n(\hat{\theta}_n)]^{-1} (\hat{\theta}_n^{1*} - \hat{\theta}_n).$$

It can be shown that the distribution of these fast and robust bootstrap estimates $\hat{\theta}_n^{R*}$ is consistent in the sense that it estimates the same limiting distribution as the sampling distribution of $\hat{\theta}_n^*$ does (see Salibian-Barrera et al. (2006), Theorem 2). Moreover, the FRB estimates $\hat{\theta}_n^{R*}$ are easy to calculate for every bootstrap sample and they inherit the robustness of the solution $\hat{\theta}_n$ for the original sample. Indeed, if an observation is downweighted in the original sample, then this observation receives the same low weight when calculating the FRB estimate of a bootstrap sample, no matter how many outliers occur in the bootstrap sample.

It has been shown by Salibian-Barrera et al. (2006) that the FRB commutes with smooth functions. In our setting this implies for instance that the sampling distribution of the coefficients $\mathbf{a} = (\mu_1 - \mu_2)\Sigma^{-1}$ of the canonical variate is estimated consistently by the distribution of $\mathbf{a}^{R*} = (\mu_1^{R*} - \mu_2^{R*}) \left(\Sigma^{R*}\right)^{-1}$. Based on the FRB distribution of these coefficients we can for example examine which variables contribute significantly to the discrimination of the two groups.

4 Applications in discriminant analysis

Based on the FRB distribution of the coefficients of the canonical variate, we can construct confidence intervals for each of the coefficients in the canonical variate or we can perform a hypothesis test to check whether a variable contributes significantly to the discrimination of the groups. This FRB test for significance of the canonical variate coefficients, can be used in variable selection procedures for robust discriminant analysis. To illustrate this, we consider the Biting Flies data taken from Johnson and Wichern (2002). The data set consists of two groups of 35 flies (Leptoconops torrens and Leptoconops carteri) and we consider the measurements `wing length`, `wing width`, `third palp length`, `third palp width`, and `fourth palp length`. The variable `wing width` contains a clear outlier in the second group as can be seen in Figure 1. Hence, a robust discriminant analysis is advisable to reduce the possible effect of outliers.

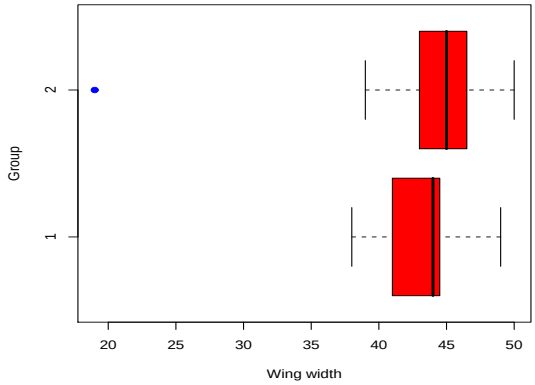

Fig. 1 Boxplot of wing width for both groups of biting flies.

Figure 2 shows the FRB distribution of each of the robust estimates of the canonical variate coefficients based on simultaneous two-sample MM-estimates of the two locations and common scatter. The solid vertical line in these histograms is drawn at the null hypothesis that the coefficient equals zero while the dashed vertical lines indicate the 95% confidence interval based on the FRB distribution. This plot already suggests that variables 3 and 5 contain the most discriminatory power, but the other variables are less relevant. If we now use backward elimination where we each time remove the least significant variable if its p-value, as estimated by FRB, is smaller than 5%, we get the series of models shown in Table 1. From this table, we see that the final model (last line) indeed only contains variables 3 and 5.

Table 1 Estimated p-values for testing, based on FRB, whether each canonical variate coefficient equals zero. In each step the least significant variable is removed if its p-value exceeds 0.05.

	Variable				
Model	1	2	3	4	5
1	0.490	0.817	0.006	0.296	0.002
2	0.306	-	0.016	0.216	0.000
3	-	-	0.016	0.096	0.000
4	-	-	0.006	-	0.000

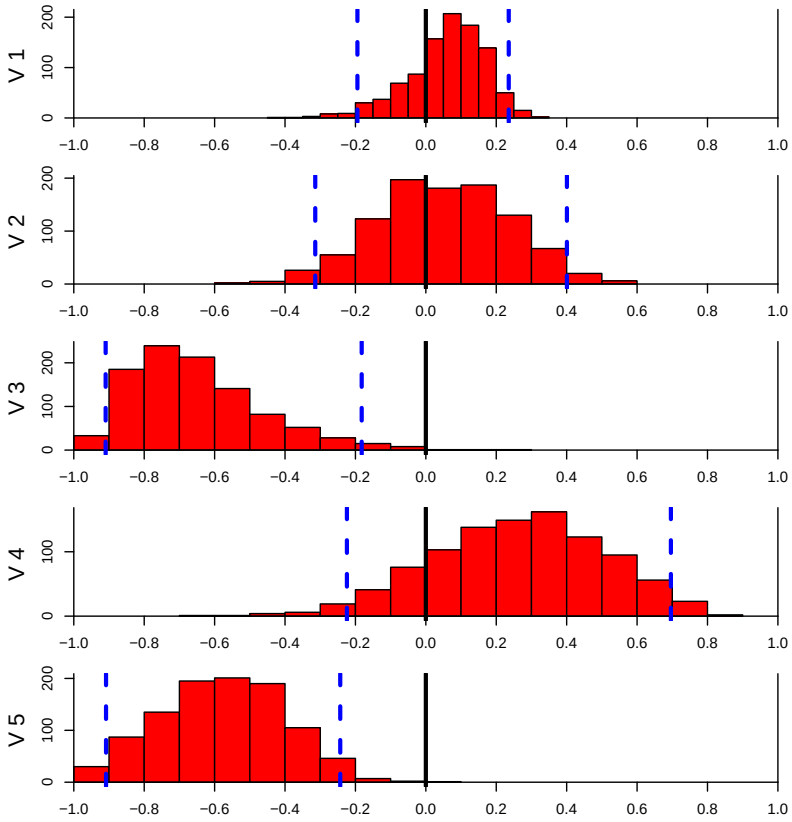

Fig. 2 FRB distribution of each of the robust estimates of the standardized coefficients of the canonical variate in the Biting Flies data.

5 Conclusion

We considered robust discriminant analysis based on robust estimates of the group centers and joint scatter matrix. We used the robust S and MM-estimates of multivariate location and scatter. The common scatter estimate can then be obtained by pooling the individual group robust scatter estimates, or alternatively simultaneous robust estimators for the locations and joint scatter estimator can be defined directly. In both cases the fast and robust bootstrap method can be used to obtain inference for the robustly estimated canonical variate. More particularly, we illustrated that the FRB can be used to construct confidence intervals for the contribution of each variable to the canonical variate and to test whether variables contribute significantly to a robust two-group discriminant analysis. A more detailed treatment and investigation can be found in Van Aelst and Willems (2010). Moreover, we illustrated that this FRB based test can be used in variable selection procedures for robust dis-

criminant analysis. We considered the two-group discriminant analysis problem in this paper, but the method can straightforwardly be extended to discrimination problems with more than two groups, by using a robust version of Wilks Lambda as in Todorov (2007).

Acknowledgements The research of Stefan Van Aelst was supported by a grant of the Fund for Scientific Research-Flanders (FWO- Vlaanderen) and by IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy).

References

1. Bianco, A., Boente, G., Pires, A.M., Rodrigues, I.M.: Robust discrimination under a hierarchy on the scatter matrices. J. Multivariate. Anal. **99**, 1332-1357 (2008)
2. Croux, C., Filzmoser, P., Joossens, K.: Classification efficiencies for robust linear discriminant analysis. Stat. Sinica **18**, 581–599 (2008)
3. He, X., Fung, W.K.: High breakdown estimation for multiple populations with applications to discriminant analysis. J. Multivariate Anal. **72**, 151–162 (2000)
4. Hubert, M., Rousseeuw, P.J., Van Aelst, S.: High-breakdown robust multivariate methods. Stat. Sci. **23**, 92-119 (2008)
5. Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis. Prentice Hall, Upper Saddle River (2002)
6. Roelant, E., Van Aelst, S., Croux, C.: Multivariate generalized S-estimators. J. Multivariate Anal. **100**, 876-887 (2009)
7. Salibian-Barrera, M., Van Aelst, S.: Robust model selection using fast and robust bootstrap. Comput. Stat. Data An. **52**, 5121–5135 (2008)
8. Salibian-Barrera, M., Van Aelst, S., Willems, G.: PCA based on multivariate MM-estimators with fast and robust bootstrap. J. Am. Stat. Assoc. **101**, 1198–1211 (2006)
9. Salibian-Barrera, M., Van Aelst, S., Willems, G.: Fast and robust bootstrap. Stat. Methods Appl. **17**, 41–71 (2008)
10. Salibian-Barrera, M., Zamar, R.H.: Bootstrapping robust estimates of regression. An. Stat. **30**, 556–582 (2002)
11. Tatsuoka, K.S., Tyler, D.E.: The uniqueness of S and M-functionals under non-elliptical distributions. An. Stat. **28**, 1219–1243 (2000)
12. Van Aelst, S., Willems, G.: Multivariate regression S-estimators for robust estimation and inference. Stat. Sinica **15**, 981–1001 (2005)
13. Van Aelst, S., Willems, G.: Inference for robust canonical variate analysis. Adv. Data An. Classif., tentatively accepted (2010)
14. Todorov, V.: Robust selection of variables in linear discriminant analysis. Stat. Methods Appl. **15**, 395-407 (2007)